The English IEP vocabulary test
Using synonyms to test vocabulary knowledge

Marlot Eline Buis
3989801
July 2017

Utrecht University – Faculty of Humanities
Master's programme Multilingualism and Language Acquisition
Supervisor: Dr. Deborah Cole
Second supervisor: Prof. dr. René Kager

Table of Contents

Abstract

An overview of the English IEP vocabulary test is presented and analysed with respect to the concepts of validity and reliability. Attention is paid to the use of synonyms in vocabulary testing, the influence of loanwords and cognates, and the question item format. All these topics were investigated and the results compared and contrasted with the English IEP vocabulary test. A recommendation is made to improve vocabulary tests to use other dimensions of vocabulary knowledge, such as pragmatic competences, to test a higher level of language proficiency. The study concludes that vocabulary test scores may be more suitable for formative assessment rather than summative assessment.

Key words; Vocabulary testing, English IEP test, primary education

1. Introduction

1.1 English in primary education and early foreign language education

English as a foreign language has been a compulsory element in primary schools in the Netherlands since 1986 (SLO, 2015, p. 7). In the Netherlands, English in primary schools is referred to as Eibo (*Engels in het basisonderwijs* / English in primary education). English education starts in groups 7 and 8 (cf. year 5 and 6) at primary school. The time allocated for English foreign language teaching (EFL) is one lesson a week, consisting of 56 minutes on average, for 36 weeks a year (Geurts & Hemker, 2013, p. 40). However, there are other English foreign language teaching methods besides Eibo in the Netherlands. These methods will be listed.

Early English education in the Netherlands has grown copiously in recent years. These vvto (*vroeg vreemde talen onderwijs* / early foreign language education) schools offer early second language education, of which English is the most popular second language. Depending on the home environment, pupils are exposed to English and Dutch outside the school environment as well. The vvto primary schools differ from Eibo schools in the respect that English education starts from group one (cf. four years old) and lasts until group eight (cf. twelve years old). Vvto schools offer a maximum of four hours of English education a week, for 36 weeks a year. Occasionally vvto schools are mistaken for bilingual education schools (*tweetalig primair onderwijs*, tpo / bilingual primary education) when no attention is paid to the fine differences. Tpo schools differ from vvto schools in the respect that tpo offer a minimum of 30 % to 50 % of all classes in English. There are currently more than 1150 schools which offer vvto in the Netherlands, of which only 50 have reached the official quality standard set by the government (Nuffic, 2015, p. 3). Only 19 primary schools currently offer tpo. These schools are part of a pilot by Nuffic, an organisation promoting internationalization in the education system in the Netherlands. The pilot was set up to aid in

the need for more English-orientated education at primary schools (Nuffic, 2017). The pilot runs from 2014 until 2019, after which the education method will be evaluated (Nuffic, 2017). An overview of the different primary education forms is given in table 1.

Table 1. Different primary education forms in the Netherlands.

	Eibo	Early Eibo	vvto	tpo
Onset second language teaching [1]& exposure	Age 9 up to 13	Age 7 up to 13	Age 4 up to 13	Age 4 up to 13
Average number of foreign language education hours per week	1 hour a week, for 2 years	1 hours a week, for 4 years	4 hours a week, for 8 years	12 hours a week*, for 8 years *including other subjects in the foreign language as well

The proficiency levels of English reached by pupils of vvto and eibo schools have been difficult to measure since no adequate measuring instrument exists. While a mandatory conclusive end test is conducted in the final year of primary school, this test does not include English competences. For this reason, a test was created by Bureau ICE, commissioned by Nuffic, to gauge the level of English after eight years of vvto education and to compare this to the results of eibo education using the same test. Between May 2016 and February 2017, over 1500 pupils have taken the test called the 'ICE Engels eindevaluatie primair onderwijs' (*IEP* /

[1] The age of onset may differ per child, as some children start primary education before the age of 5. This results in the possibility that a child of 10 and 9 years old have had the same amount of EFLT, despite the one year age difference.

ICE English end evaluation primary education) (Bureau ICE, personal communication, March, 2017). The test may be given at the end of primary school in the Netherlands, and consists of 45 vocabulary questions, 15 listening questions and 15 reading questions. The results are presented in the form of the Common European Framework of Reference (CEFR), level A1, A2 or B1. The ICE English end evaluation primary education test will be referred to as the English IEP.

1.2 Mandatory conclusive test

In primary schools in the Netherlands children, are tested in several subjects such as geography, history, and Dutch vocabulary. A final conclusive test has been made mandatory for the eighth and final year of primary school by the Dutch government in 2015 (Toetsbesluit PO, 2014). The choice lies with the school as to which test the pupils will take. The results from whatever test is selected, together with an evaluation from a pupil's primary school teacher, are used to decide which form of secondary education the pupil should be enrolled in. Thus, these tests often have far-reaching implications. The three forms of secondary education in the Netherlands are preparatory secondary vocational education (*VMBO*), senior general secondary education (*HAVO*), and university preparatory education (*VWO*) (Nuffic 2015). These final conclusive tests often have items concerning language, math, and world orientation (CITO, N.D.). While there are a few possible choices regarding conclusive end tests, one of these is used more than others. This widely used tests has been created by CITO (*Centraal Instituut voor Toetsontwikkeling* / Central Institute for Test development). The CITO test is a popular test to measure pupils' knowledge after eight years of primary school education mainly because until recently there was little to no choice in conclusive end tests. This test is supposed to give a valid recommendation concerning the pupils' current and future competences (CITO, N.D.). However, since tests given at the conclusion of primary

education, such as the CITO, have become mandatory, criticism regarding the vocabulary part of these tests has increased.

There are several reasons why vocabulary is such an important competence in primary school, secondary schools and beyond. First of all, vocabulary plays a central role in reading competences (Stahl & Nagy, 2006; Hattie, 2009). Moreover, vocabulary is necessary for speaking and writing competences (Nation & Snowling, 2004). A broad vocabulary ensures children are able to comprehend new information offered in classes, and also aids children in making new connections faster (Marzano, 2013).

Vocabulary became a popular topic in the 1990s with scholars interested in the conceptualisation of vocabulary knowledge and how to measure this (Read, 2000). The knowledge of vocabulary was mainly discussed in terms of dimensions. There are three types of vocabulary dimensions; the dimension of size, which refers to the number of words an individual knows, the dimension of depth, referring to how well the words are known, and the receptive-productive dimension, which is the relationship between words which are recognised and words which are used (Nizonkiza, 2016). All three vocabulary sizes will be analysed by the author of this thesis, for the purpose of gaining a deeper understanding of the current vocabulary knowledge. While the focus of researchers used to be placed on vocabulary size tests (eg. Nation (1990) attempted to validate vocabulary size tests (e.g. Vocabulary Levels Test), the focus has shifted towards other dimensions of vocabulary testing over the years. While the focus of researchers lay mainly on English vocabulary, some of these vocabulary tests were translated into other languages by researchers and educators. Depth vocabulary tests were created (Read's (1993) Word Associates Test), as well as productive vocabulary tests such as Laufer and Nation's (1995) Lexical Frequency Profile. The depth vocabulary test was used to measure how well vocabulary items are known. The degree to which an individual has mastered vocabulary words is an important factor because

this results in how they are able to employ these words, receptively or productively, and appropriately. This is tested through questions with a multiple-choice format, inquiring about words their synonyms and collocates. The results of the test are used to calculate the size of the vocabulary of the test taker. The words in the vocabulary size test are categorized on word frequency. The answers of the test taker and the word frequency are used to estimate vocabulary size by using a weighted average. Productive vocabulary tests, such as Laufer and Nation's Lexical Frequency Profile (1995), measure productive abilities by offering controlled-production questions, such as: "The garden was full of fra__flowers" (Laufer & Nation, 1995). This method offers an incentive for the correct word by providing the first few letters, which is called cloze testing, but has its drawbacks in that it only allows for one context-correct word. Furthermore, the terms receptive and productive knowledge are based on a categorical difference between different definitions of what it means to know a word. Different elements of a word may already be known, while others have not yet been acquired. On a word level knowledge consists of phonological, orthographic, morphological, syntactic and semantic knowledge while on a broader level it contains the ability to be able to actively recall words in colloquial use and while writing. Other aspects of word knowledge include pragmatic use, formality, connotations, and cultural awareness of sensitive words. Nation notes that at the most basic level knowing a word includes form, meaning and function (1990). To illustrate this, receptive word knowledge of the word 'impartial' would include being able to have mastered the following features;

- to recognise the word upon hearing it (form)
- to recognise the written form of the word (form)
- to recognise the adjective *partial* and the prefix *im-* (form)
- to recognise the connotations the word signals (meaning)
- to recognise when the word is used correctly and incorrectly (meaning)

- the collocations with which a word typically occurs (use)

- knowing whether it is a common word or a pejorative (use)

(Nation, 1990, 27).

Productive word knowledge includes a different array of competences and knowledge. Productive knowledge of the word 'impartial' would include the ability;

- to be able to use the word correctly in a sentence (use)

- to be able to pronounce it correctly including the word stress (use)

- to be able to write it correctly (form)

- to be able to clarify the meaning of 'impartial' with words (meaning)

- to be able to use it in different contexts expressing the different meanings of the word (meaning)

- to be able to use the word depending on the formality of the context (use)

- to be able to produce synonyms and antonyms for 'impartial' and being aware of the subtle differences between them (meaning)

(Nation, 1990, 28).

The many different aspects that make up knowledge of a word should not be seen as an all-or-nothing dichotomy, but rather as a scale of knowledge where one aspect may be acquired and another is yet to come (Pavičić Takač, 2008).

Even though all the dimensions of vocabulary knowledge are testable, the most often used vocabulary tests continue to be the size test and the receptive vocabulary knowledge test. Several studies have shown a strong connection between receptive and productive vocabulary knowledge and (future) reading and writing competences (Staehr, 2008; Webb, 2009). Furthermore, a predictive relationship was established between receptive vocabulary knowledge and linguistic proficiency (Beglar, 2010, Meara, 1996, Meara & Buxton, 1987, Meara and Jones, 1988, Nation, 1990). While understanding a word aids in reading ability and

being able to use a word facilitates in writing ability, the underlying nature of the relation between the vocabulary test results and the reading and writing competences remains unclear and little to no research towards this has been done. Furthermore, it is unclear whether all different vocabulary tests and the question formats they employ result in comparable predictability properties. Thus, while the nature of the relationship between receptive vocabulary and linguistic proficiency is unclear, the predictability properties of the vocabulary test have been empirically proven to exist. However, while the reasons for testing vocabulary are clear, the motivation for choosing the question item formats for vocabulary testing are not.

1.3 Vocabulary test and dependent and independent models

The test specialists who develop vocabulary tests focus on two types of methods; method-dependent vocabulary tests[2], and method-independent vocabulary tests (e.g. the CITO vocabulary test). A method-dependent vocabulary test has extracted the words which are tested from the method and material offered in schools. Compared to a method-independent vocabulary test which is independent of the method and materials offered in school and focuses on words which the pupil has had a high chance of encountering. The method-dependent and method-independent vocabulary test can be constructed to measure either one of the three vocabulary dimensions (depth, size, and receptive-productive abilities). More often than not, the decision for choosing any particular question format when constructing a test is made on a practical or traditional basis instead of on empirically grounded rationales (Kremmel & Schmitt, 2016). A practical consideration when constructing a vocabulary test would be to opt for a multiple-choice question format rather than an open answer one, which would save time grading the tests.

[2] Method-depend vocabulary tests may be construed upon an education method. Thus these tests may differ among each other if they are constructed upon different educational methods,

The method-dependent vocabulary test is based on the teaching method used in a specific primary or secondary school. A teacher instead of a test construction specialist will often make a method-dependent vocabulary test based on the words used in class and throughout the method. The words the pupils encounter in their books and during classes at school are thus tested using a method-dependent vocabulary test. Method-dependent vocabulary tests will often be used to measure short term vocabulary acquisition, and are able to compare pupils per class because all the pupils are offered the same vocabulary through the method and hear the same words in class. Although method-dependent vocabulary tests may be constructed to test one of the three dimensions of vocabulary (size, depth and receptive-productive abilities), they are often used to test vocabulary size and receptive vocabulary knowledge, mainly due to practical reasons such as an easier construction and control of the test, and to more easily compare pupils with their classmates and earlier achieved results (Kremmel & Schmitt, 2016).

In contrast to this, method-independent vocabulary tests are not based on one specific teaching method, but are often based on frequent words pupils will have a high likelihood of encountering during school. Method-independent vocabulary tests are often used to compare pupils on a nationwide basis, which is possible due to the large numbers of pupils who take a method-independent test and the fact that these tests are standardised. These tests use a multiple-choice question format, and by inquiring about words which on average should have been encountered previous to the moment of testing. A schematic overview of a few examples of method-independent vocabulary tests [3] are given in table 2.

[3] The creator of the test is between parenthesis. The test is graded by the educator.

Table 2. Schematic overview of method-independent vocabulary tests per dimension

Dimension	Method-independent vocabulary test	Receptive	Productive
Size	Vocabulary levels test (Read, 1990)	Yes	No
Depth	Word Associates Test (Read, 1993)	Yes	No
Receptive-productive	Lexical Frequency Profile (Laufer & Nation, 1995)	Yes	Yes

Thus, while all dimensions of vocabulary are able to be measured through method-independent vocabulary testing, the most often opted choice is testing the vocabulary size and vocabulary receptive knowledge tests due to practical reasons such as the time constraints open questions pose for grading. However, the method-independent vocabulary size tests' results may be interpreted wrong, because while they are unsuitable to measure short term results in terms of how many new words a pupil has acquired since the last vocabulary test, they are often used for this purpose (Schmitt, 2000; Beck, McKeown, Kucan, 2008). Short term progress is best measured with method-dependent vocabulary test, which is able to measure the number of words which are known from the amount of vocabulary words offered by the teacher and throughout the method used in class. When using method-dependent vocabulary tests, the words are extracted from the method offered in school and the material offered in class. Furthermore, the use of high frequency words and low frequency words can be endlessly alternated in a vocabulary test, but the results of a vocabulary test are often constrained by variables such as the personal environment of the test taker, the educational method offered in school, and the test taker's personal aptitude of for language. Moreover, method-independent vocabulary size tests may have some psychometric problems as well. When the vocabulary size of a (foreign) language learner is calculated by extrapolating a number from the results of a vocabulary test, even a test with high frequency words, the results become statistically unreliable (Browne, Chichi & Culligan, 2007). The results (the

estimated size of the vocabulary) are based on previously collected data (the vocabulary test), and the conclusions cannot be fully hedged since previous collected data may not resemble future data (the real size of the vocabulary). The extrapolation of the vocabulary size from a test is not a preferable option to measure vocabulary size, while testing all supposedly known words is unrealistic as well, due to time and effort constraints from both the teacher and the pupil.

Thus, the problematic areas of method-independent vocabulary size tests lay in the disregard for personally acquired vocabulary, and the difference in materials and methods offered in schools. In addition to this, the extrapolation of results to calculate the size of a pupils' vocabulary is also problematic. These factors may ensue in the use and construction of unreliable method-independent vocabulary size tests which are often used in the conclusive end tests at primary schools.

The problems when using method-dependent vocabulary tests do not include any psychometric problems, as long as the goal is to test whether the vocabulary offered in class and throughout the education method is known by the pupil. Thus, method-dependent vocabulary tests are more attuned to the personal school environment of the pupils, because the words are extracted from the material offered to the pupils. Nevertheless, method-dependent vocabulary tests are unable to compare pupils on a nationwide basis, and do not take into account the exposure to foreign words pupils may encounter outside of the educational environment, making this method unsuitable to include in a conclusive final test at primary school.

Not only have vocabulary tests been criticised on whether they are method dependent or independent, but the vocabulary test items have been criticised on their own as well. It is unclear what competences are being tested precisely; whether it is memorised words, partial knowledge of a word, or the ability to be able to guess the correct answer on a multiple-choice

question. However, the problems concerning vocabulary testing are not restricted to the vocabulary part of conclusive final end tests. Almost all vocabulary tests may produce invalid results in some way.

Overall vocabulary tests have several problems and a range of these problems concerning vocabulary tests has been discussed, revealing three main criticisms: What competences do these vocabulary tests 'test'. Whether vocabulary size tests measure receptive (recognizing the word) or productive (being able to use the word) knowledge has not been investigated thoroughly by researchers; more often than not, both the method-independent vocabulary size test and the method-dependent vocabulary size test are considered to test receptive vocabulary knowledge, thus vocabulary which is needed for reading and listening. However, no empirical evidence to this effect has been found (Kremmel & Smith, 2016). Another problem that arises when analysing vocabulary tests is how reliable the results are. If the competences that are being tested are unclear it remains a challenge to interpret the results of a vocabulary test in a valid and reliable manner. Furthermore, the relationship between the predictability properties and whether this is the same for receptive and productive competences in all test formats remains unclear.

It is necessary to understand what competences are being tested before being able to interpret the results in a reliable manner. Based on these three problems, the following research question will be addressed in this study:

What competences are being tested with vocabulary test items?

This study also seeks to address the following sub-questions:

1) What are the predictability properties of vocabulary test items for Dutch English foreign language learning children and their future English competences?

2) How reliable are vocabulary test items?

3) How does the reliability differ among vocabulary question item formats?

4) In what way can improvements be made to the English IEP vocabulary test regarding reliability, and predictability properties?

The preceding questions will be investigated and the results will be compared and contrasted with the current English vocabulary IEP test. The answers to these research questions will be based on research towards the English IEP vocabulary test, but may be used to discuss other similar vocabulary test. Furthermore, an overview of related work in the academic field will be presented.

2. Theoretical Framework

2.1 English IEP test for Dutch learners of English as foreign language

The English IEP was created by Bureau ICE, in commission by Nuffic, to gauge the level of English at the end of primary school. The test may be used to measure the results of different types of English educational methods by comparing the results of the test on the common European Framework reference (CEFR). An explanation of the CEFR is given on page 17. The test may be given at the conclusion of primary schools in the Netherlands. The English IEP is an adaptive test and contains 45 vocabulary questions, 15 listening questions and 15 reading questions. The vocabulary test is in British English, since this is taught in the majority of primary schools. However, no productive vocabulary is tested, which makes the English IEP vocabulary test also suitable for those who have learned American English in primary school. Even though pupils who have learned American English may come across differently spelled words in the English IEP vocabulary test such as *recognise* instead of *recognize*, this may not pose any major problems for pupils since the orthographic similarities between the words surpasses the differences. The test is divided into two parts, the vocabulary and listening part (part A), for which pupils without learning disabilities such as dyslexia have

a set maximum time of 45 minutes, and the listening and reading part (part B), which also has a set maximum time of 45 minutes. Based on their results on the vocabulary part of the test, the pupils are either directed to an A1/A2 level or an A2/B1 level for the listening and reading part. When the test taker has finished part B, the results are presented in the form of the Common European Framework of Reference (CEFR), e.g. level A1, A2 or B1. More information on the CEFR may be found in 2.2, on page 16.

The vocabulary part of the IEP test is the same for every pupil: the format of the items, the different CEFR level of the words, as well as the amount of nouns, verbs and adjectives occurring in the test. The vocabulary part of the IEP contains 15 questions on A1, 15 questions on A2 and 15 questions on B1 CEFR level. The distribution between nouns, verbs and adjectives and adverbs has been made based on the frequency with which these words appear in the 'real world'. Nouns and verbs are both represented for 40 % each, while adjectives and adverbs are both represented for 20 %. This distribution is the same for each CEFR level.

The different question item formats have also been carefully distributed among the different vocabulary tests by the test makers of the English IEP vocabulary test. The multiple-choice format is on average 15-30 % of the test Multiple matching makes up 10-25 % on average of the total test, while the mix and match makes up 10-15 % of the test. Other formats which are used in the English IEP vocabulary test are words and picture, which makes up 15-30 % of the test, while picture and words, makes up 10-30 %. The difference between these testing formats lies in the manner in which the question is construed and the possible answers a test taker can choose from. The percentages given above are not absolute numbers. How often a question item format occurs depends on the CEFR level of the questions. For example, when the items are on the A1 level on the CEFR, there are 3 multiple-choice items, compared to the A2 level, where there are 2 items in this format. The items are always distributed

according to the previously given percentages, but can as illustrated differ slightly per CEFR level.

2.2 Common European framework of reference

The CEFR was designed to create a common framework of foreign language proficiency throughout Europe. Furthermore, it provides "a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks etc. across Europe" (Council of Europe, 2001, p. 1). It covers three areas of language knowledge; reading, hearing, and writing competences. The CEFR has been widely accepted throughout Europe and has been implemented in many schools, exams, and course books. Several CEFR guides have been developed for a number of European languages such as French, German, and English. The CEFR describes what language learners need to be able to do to use a language efficiently at each level in 'can do' statements. These statements were developed by the Association of Language Testers in Europe; for a sample of this see Appendix A, on page 80.

The CEFR includes three different levels of proficiency, which are further divided into two components per level. The first level is A, consisting of an A1 level and an A2 level and covering the basic use of a language. The second level is B, which consists of B1 and B2. These levels describe more independent users of a language. The third level, C, which also has a C1 and C2 level, describes proficient language users. For an overview of the levels and their users see figure 1.

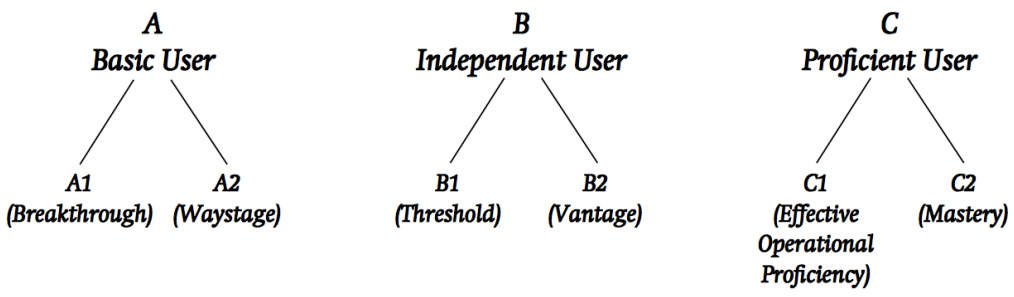

Figure 1. CEFR levels (Council of Europe, 2001)

The CEFR has been implemented in many schools in the Netherlands, where it is currently used to set goals for foreign language learning in both primary and secondary schools. The central conclusive language exams at the end of secondary school have also been linked to the CEFR to demonstrate which pupils achieved a certain CEFR level. However, the CEFR has not (yet) become a compulsory component of education in the Netherlands (ERK, N.D).

Two problems concerning the CEFR may be that the framework is too vague and open to interpretation. An example of this may be that the descriptions per level are multi-interpretable. While this may intuitively seem to be a positive point, it may result in an unbalanced framework if not all countries are able to interpret the descriptions in a somewhat similar manner. The second problem is that attitudes towards competences in certain foreign languages may also differ. A teacher in Germany and a teacher in the Netherlands may both use the term *simple*, while each employs an entirely different concept of the word. De Saussure (1916) explains this as a difference between the signifier (sound pattern of the word) and the signified (the concept). As long as the term *simple* is not further elaborated upon, such differences between respective language users may result in two different working terms of the word *simple*. Many different factors influence what words or constructions are viewed as simple; the language family the foreign language belongs to, and whether this differs from the native language family, cognates, loanwords, but also media exposure in the respective language in the country. For example, Lotto and de Groot (1998, p. 32) found that learning words which are cognates is easier than learning words which are not. While the cognate status of words might not directly influence the evaluation process of what constitutes simple, it does influence the difficulty level of the learning process and thus the difficulty level of the vocabulary word (Lotto & de Groot, 1998, p. 32; de Groot & Keijzer, 2000).

3. Method

3.1 Vocabulary and memory

Even though vocabulary acquisition has been studied extensively for several decades, a generally accepted theory of vocabulary acquisition does not yet exist, and a possible explanation for this may be lack of cooperation or disagreement between experts (Pavičić Takač, 2008). However, several popular theories on foreign language vocabulary acquisition, and the relation between vocabulary and memory will be discussed in the following chapter. Possible implications from this research will be drawn.

The influence of the L1 on learning vocabulary in L2 is dependent on several factors. Since an L2 learner has already acquired the concept of the word and its associated meaning in the L1, L2 vocabulary acquisition differs from L1 (Pavičić Takač, 2008). A few factors may influence L2 vocabulary acquisition: whether the L1 and L2 are related, if the L2 occurs often in the cultural or social environment of the L1, and the attitude in the social environment of the L1 towards the L2. These factors all influence how often and where a pupil may come across written or spoken words or phrases, and thus in part how likely an L2 learner is to acquire a large vocabulary. Furthermore, it will influence how many L2 cognates and loanwords there will be available in the L1. Even though the input in the school environment may be just as crucial, factors such as the cultural and social environment of the L1 should not be disregarded when investigating vocabulary acquisition in the L2. While L2 learning is often limited to classroom exposure, in the Netherlands L2 learners are often exposed to English outside of the classroom as well. English programs are broadcasted on Dutch television, learners play video games in English, and listen to English music (Verspoor, 2010). Furthermore, with a close proximity to the U.K., and a globalizing media, English is often heard or seen in advertisements (Gerritsen, Van Meurs & Gijsbers, 2000).

Vocabulary testing is often regarded as retrieving the phonological representation and meaning of a word from memory (e.g., Ehri & Rosenthal, 2007). While research towards vocabulary testing is a relatively new domain in empirical research, the mechanisms of learning new vocabulary have been discussed thoroughly. The newly acquired vocabulary words are stored in the *long term memory* (LTM), and new information enters into the *short term memory* (STM) (Atkinson & Shiffrin, 1968). This model of the memory components is an early and well-accepted model, to which Baddeley and Hitch have suggested adding a *working memory* (WM) unit (1974). The WM is responsible for retrieving information from the LTM and from the outside world. The WM is active in learning, comprehension and reasoning (Baddeley, 2003). In this revised model by Baddeley (2003), the WM includes four components: a visuo-spatial sketchpad (VSS), the episodic buffer (EB), the central executive (CE), and the phonological loop (PL). The VSS provides storage for a limited amount of time for visual and spatial representations. The EB integrates elements from the LTM and the WM into a 'multi-dimensional code' (Gathercole & Alloway, 2008). The EB links information across domains to form integrated units of visual, spatial, and phonological information with time sequencing such as the memory of a story or a movie. While the PL offers temporary storage for phonological information, the CE controls the attention and regulates the flow of information between LTM and WM systems, as well as within WM (Vulchanova, Foyn, Nilsen, Sigmundsson, 2014). Both the STM and the WM have been shown to be active while learning new vocabulary words. The memory store should be regarded as a dynamic organizational structure, wherein new information is assimilated into the old (Paivio, 2007). The interplay between these components may work in a similar manner when retrieving words from LTM, however this has yet to be empirically proven.

Ricketts, Bishop and Nation found that the role of orthography in vocabulary learning facilitates the retrieval of new words (2009). They found that when children who are learning

new words are incidentally exposed to their spelling, these children are able to more easily retrieve the new words compared to children who were not exposed to the spelling of the new words (Ricketts, Bishop & Nation, 2009). Li, Zhang, Ehri, Chen, Ruan and Dong call this easier retrieval of a word after being exposed to the spelling (form) orthographic facilitation (2016). One theory which explains the function of orthographic facilitation is a connectionist theory by Ehri (1992). According to Ehri's amalgamation theory the process of learning new words involves forming new connections in the brain, called orthographic mapping. The graphemes (smallest unit of written language) of specific words are mapped onto phonemes in the pronunciation, alongside the meaning of the word, and stored in memory – this mapping into one memory unit is called an amalgam (Li et al., 2016). The idea that memory is not a single unit, but rather a system containing separate but interacting components has been accepted for quite some time, though the specific workings remain hard to describe (Baddeley & Hitch, 1974). A similar explanation has been given by Perfetti's lexical quality hypothesis (2007). The lexical quality hypothesis states that when the orthography is mapped onto the other identities of the word (phonological, meaning) this enhances the quality of the representation compared to words lacking this representation. Furthermore, a word which has a mapped orthography is easier, more accurate and more efficient to retrieve compared to a word which lacks an orthographic component (Perfetti's, 2007). To retrieve information after it has been mapped, the central executive storage, which is a domain-general component, is responsible for the retrieval of long term stored information; but also for controlling and monitoring information and attentional control (Baddeley & Logie, 1999). Thus, it may be argued that when more components are mapped onto a word, the word will become easier to retrieve.

For a visual representation of the memory unit and how vocabulary words are stored, see figure 2.

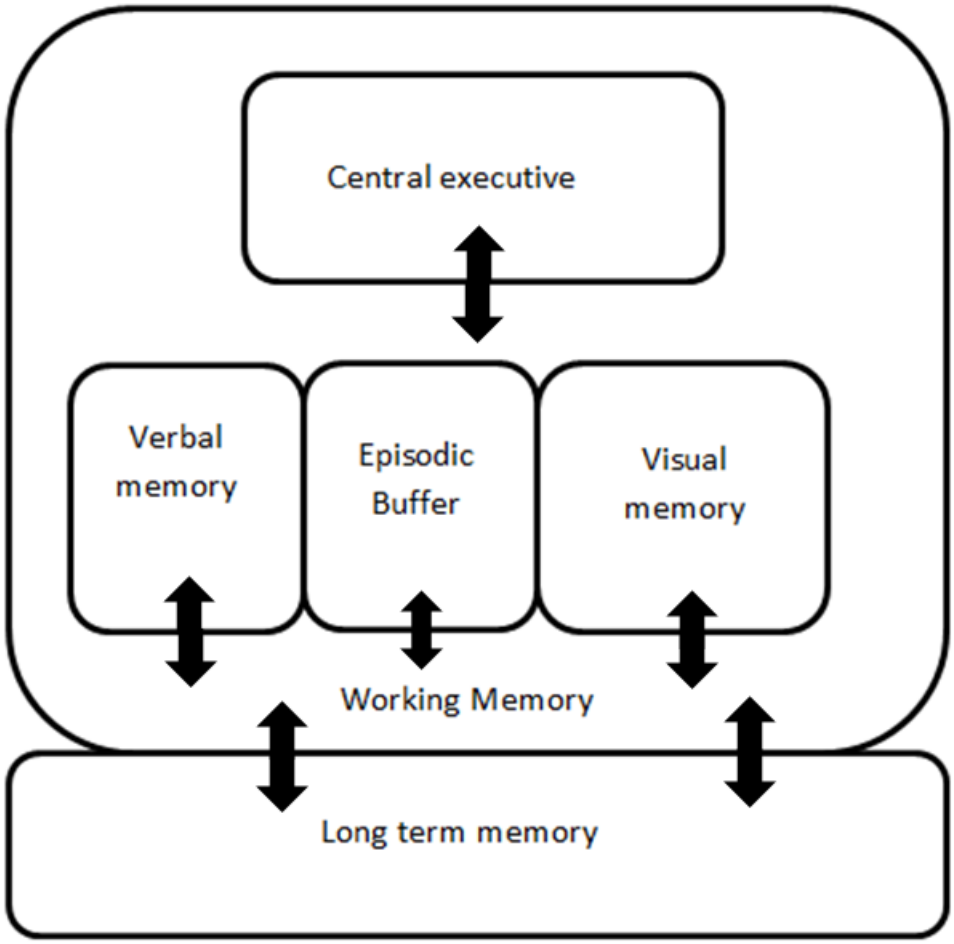

Figure 2. Visual representations of memory units (Buis, 2017).

Several factors influence the learning of new vocabulary words, such as phonological (verbal) or visual information which has been mapped upon the word during the learning process. The influence of being exposed to the word form, or the spelling has an empirically proven impact on learning new vocabulary words. When 10 year olds were exposed to the spelling of words during learning trials, word pronunciation and meaning became easier to retrieve than if they had not been exposed to the word pronunciation and meaning (Li et al., 2016). The exposure to word form (orthographical) may also be referred to as visual (denoting form). Here the terms are used interchangeably. Several studies found indications for confusion in memory for visually similar data, such as words which were visually similar (e.g. *were-where*) (Logie, 2014). When a subject attempts to remember visually presented letters or

characters, confusion about the correct form may be observed, as happened in the study by Hue and Erickson (1988). In their study, subjects tried to immediately recall unfamiliar Chinese characters which had a similar appearing counterpart. This resulted in a visual similarity effect (e.g. visually similar characters were often confused with each other) (Erickson, 1998). Further visual confusion errors were found in the verbal recall of visually presented stimuli (Wollford & Hollingsworth, 1974). The visuo-spatial system was found to be responsible for the orthographic facilitation.

The process of phonological (or verbal) memory aiding in vocabulary acquisition has also been the subject of many studies and evidence that phonological working memory is related to the acquisition of vocabulary (and grammar) in L1 and L2 has been empirically proven (Adams & Gathercole, 1996, 2000; French & O'briend, 2008; Masoura & Gathercole, 2005). Verhagen, Messer and Leseman (2015) found that phonological short term memory and working memory relate to the acquisition of vocabulary and grammar the same way for L1 children as for L2 children who acquire their L2 naturalistically (through submergence). This finding might be an indication that when a second language is learned naturalistically, the mapping process of information in memory follows a similar path as in the L1. Furthermore, phonological short term memory is regarded as a necessary factor for the development of stable phonological representations in long-term memory, which is needed for acquiring both vocabulary and grammar (Baddeley et al., 1998). Moreover, a positive relationship between sounds and letters was found by Ricketts, Bishop and Nation (2009). They found that orthographic facilitation was influenced by consistency between letters and sounds in the spelling of words when learning for children aged 8-9 novel words (Ricketts, Bishop & Nation 2009). The influence between word form and sounds was also found by Rastle, McCormick, Bayliss and Davis (2011). They observed this consistency effect between

letters and sounds (2011). Their study involved a speech perception and production task in which associations between novel words, pictures and spelling were learned.

The different theories all indicate some form of separate but interlinked storage units which benefit from orthographical and phonological mappings. An assumption which arises is that several vocabulary question item formats may rely more heavily on either phonological or orthographical mapping. However, it might be difficult to prove that certain tasks demand more from phonological memory than from the visuo-spatial system. When different levels of performance are achieved for phonological and visuo-spatial tests, this may reflect the demand a task has on a specific system. However, it might appear as if performance indicators are easily comparable; but a 50 % score on a phonological task might not be similar to a 50 % score on a visuo-spatial task. The only possibility to compare these two systems is if the same cognitive system underlies the performance of both these tasks and if the scoring method assesses the task in a similar way (Logie, 2014).

The relationship between long term memory and the different components in the WM and how these interrelate may furthermore differ between L1 and L2 retrieval of words. In an empirical debate which took place from 1950 up until 2017 and will most likely continue in the future, the focus lies on the contrast between single stored memory systems and multiple stored memory systems (Paivio, 1981). The theory that bilingual and second language learners' memory is made up of single storage systems (LTM) in which both L1 and L2 are stored, notes that the languages are interdependent, but not dependent from each other. The dual memory system theory notes that the systems for language memory can function independently or cooperatively in a different array of tasks, but that L1 and L2 are stored in separate storage systems (Pavio, 1981). While differences exist between the single memory storage system and the dual memory storage system, future research should focus upon one of the core issues: namely, whether L2 word information will be differently extracted from L1

pupils, compared to L1 information retrieved from LTM from L1 pupils. However, the problem remains whether these two systems emerge right from the beginning of learning an L2 or if they gradually form their separate but interlinked system, after enough input has been provided. A pupil who has had 4 or 7 years of foreign language education may already have built such a separate system, but a pupil who has just learned 10 words in a new foreign language might have not. Another possibility is that a pupil who has had 4 or 7 years of foreign language education may have built a strong second language memory system. While the pupil who has just started has a weak second memory system, depending mainly on the first memory system.

It may well be possible that a specific vocabulary question item format relies more heavily on phonological memory than on visual memory, and when the participant has a weak phonological memory, this may be apparent in the results of a test as well. While more in-depth research is necessary to provide conclusive evidence on whether specific question formats use specific memory units, the hypothesis that different vocabulary items access different memory units (phonological, visual) seems prima facie possible. However, a note with this hypothesis is that it does not account for the relation between native (L1) and second language (L2) vocabulary extraction. When the extraction of L2 words from memory is studied, caution must be applied because the mapping of graphemes, and phonemes might not be available in a similar manner as in the L1, and depending on the strength of the second language knowledge these might even be accessed through L1 memory form. The author of this study has found no previous studies that have focused on different memory tasks related to different vocabulary question items formats.

3.2 Synonyms and meaning

In vocabulary testing a popular question method is the multiple-choice method. The question contains a word and the four or six possible options contain one synonym which is the correct answer. However, do these words in fact express the same meaning on all levels, regarding the conceptual, register, language user, connotation and denotation aspect? Would there than be use for two words denoting exactly the same concept? Hayek discusses this issue regarding communication as follows:

> It would […] not be possible to discuss the phenomenal world with other people if they did not perceive this world in terms of the same, or at least of a very similar, order of qualities as we do. This means that the consciousness of other people classifies stimuli in a manner similar to that in which our own mind does so, and that the different sensory experiences are 'subjective' in the sense of belonging to the perceiving subject as distinguished from 'objective' (belonging to the perceived objects) – a distinction which is the same as that between the phenomenal and the physical order – it is yet inter-personal and not (or at least not entirely) peculiar to the individual (Hayek, 2014).

While the issue of *true* communication has been the centre of philosophical debates for a long time, the concept of testing vocabulary through synonyms has not been debated as thoroughly. Thus, a word may denote a slightly different object than its synonym in the L1, and this concept also appears in translations from L1 to L2. While often words in the foreign language may appear to be similar to those in the native language, the referents, connotations, or implicatures may be different. These differences not only arise when naming words from abstract or socially constructed domains such as emotions, but also when naming concrete nouns referring to common objects (De Groot, 1993). De Groot notes that "it is a well-known fact that complete meaning equivalence of the two terms in a translation pair is a rare

phenomenon" (2011, p. 132). While a translation of the French word *balle* will produce the English word *ball,* the French word can only be used as a referent to tennis balls, but not to denote basketballs or footballs (Paradis, 1997). The referent of a word may be different in the L1 than in the L2 translation and cause confusion if not all aspects of a word are known (such as the possible referents of the French word *balle*). Furthermore, some foreign language learners might not be aware of certain pragmatic differences between translation equivalents. It is important for researchers to understand how bilinguals and foreign language learners link words with meaning, but cross-language differences make this difficult to investigate.

A study by Malt, Sloman, Gennari, Shi and Wang (1999) revealed that when speakers of American English, Mandarin Chinese, and Argentinean Spanish had to label 60 common household containers the English participants named 16 containers *jar, bottle* and *container*, while Argentinean Spanish participants gave these 16 containers 7 different names. The Chinese participants in this study labelled 40 different objects with one label, which were labelled *jar, bottle* and *container* by the English participants. These differences illustrate the cultural dissimilarities as well as the way words are employed for each context and culture differently. In a study by Ameel, Storms, Malt and Sloman it was found that the naming pattern for common household objects between monolinguals and bilinguals was influenced by the history of their languages (2005). Ameel et al. note that a language's vocabulary is shaped by *convention, pre-emption* and *chaining* (Ameel et al., 2005*)*. For an illustration of a combination of convention and pre-emtion (A and B), and an illustration of chaining (C and

D) see illustration 1.

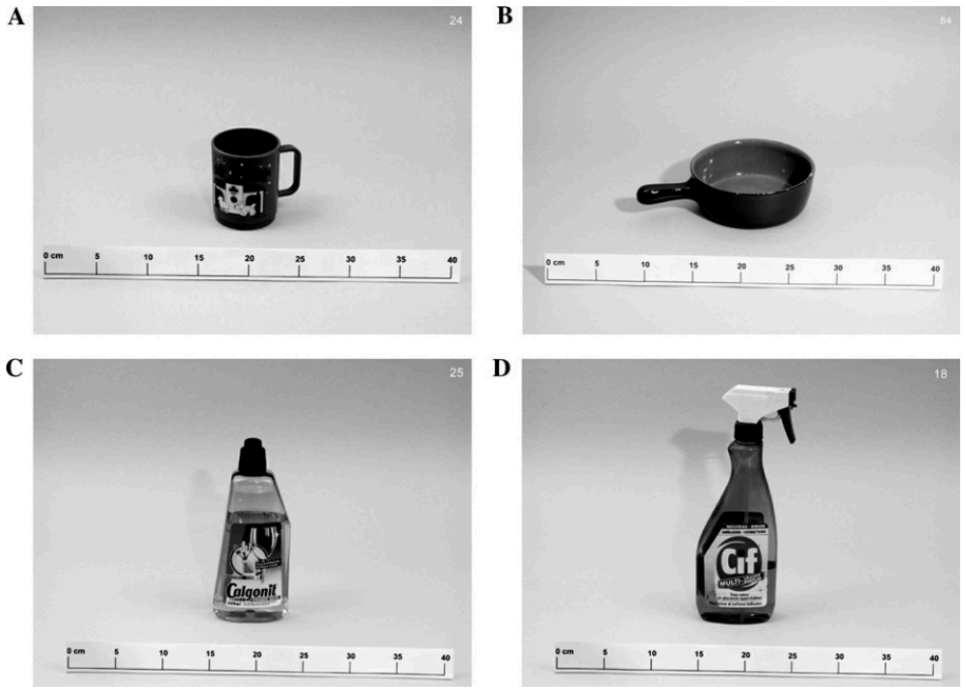

Illustration 1. Examples of *convention, pre-emption* and *chaining* (Ameel et al., 2005)

Ameel et al, describe the relations of the following objects as such: (A) is an object belonging to the dishes set named *beker* by Dutch-speaking monolinguals with higher average similarity to the *tas* category and the nearest neighbor being a *tas*. While (B) belongs to the category of the dishes set named *caquelon* by French-speaking monolinguals with higher average similarity to the *plat* category and the nearest neighbor being a *plat*. Object (C) belongs to the bottles set, named *fles* by the Dutch-speaking monolinguals, with higher average similarity to the bus category. Object (D) is an object of the bottles set named spray by French-speaking monolinguals with higher average similarity to the *bouteille* category (Ameel et al., 2005).

Certain names may be used due to "linguistic convention rather than because of specific similarity relations to other objects associated with the category name" (Ameel et al., 2005). When a particular name is not used to refer to an object to avoid confusion or ambiguity with another similar object; this is an instance of pre-emption (Ameel et al., 2005). The term chaining denotes the situation wherein two objects appear similar, but one of them is

more typically associated with the traditional name, the lesser typically appearing object may receive a different name (Ameel et al., 2005). However, while these mechanisms clarify how similar objects may be given different names, the process of producing the correct name in a situation is more complex.

Nation (1990) noted that productive knowledge of a word includes being able to produce and use synonyms as well as antonyms for a word, and this word knowledge is often tested using a multiple-choice question, while the subtle differences may be unable to be measured in this manner. Distilling a word out of context and asking someone to provide or choose the correct synonym will ensure that the differences, such as knowledge of pragmatic, or formal features, between the word and the synonym are no longer necessary information in answering the question. The true nature of word knowledge becomes, by measuring one word knowledge one dimension, lost in this aspect . The different features of a word are visually illustrated to demonstrate the differences between two manners of representation. Phonological representations of a word may be found under the verbal (phonological) representation unit. While the form of the word may be found under orthographical representation.

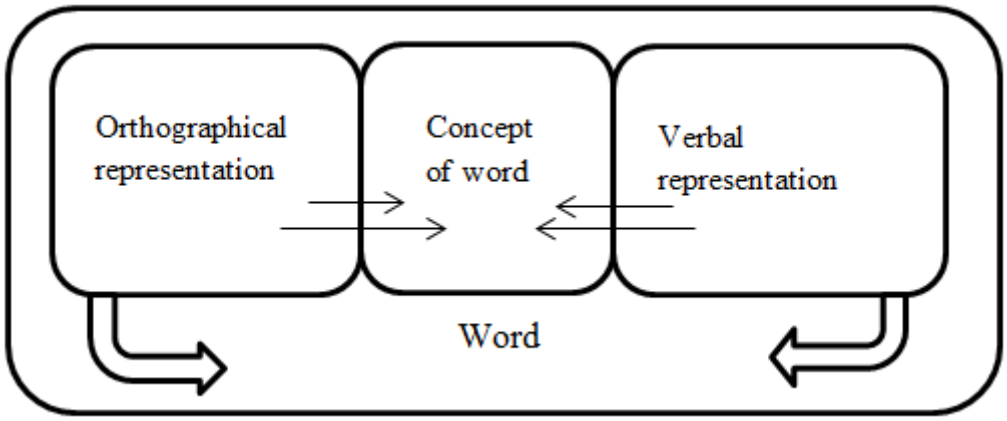

Figure 3. Visual representation of basic word knowledge (Buis, 2017).

The basic concept of vocabulary knowledge is represented in figure 3, while the reality may be a bit more complex, and a possible representation of this is given in figure 4.

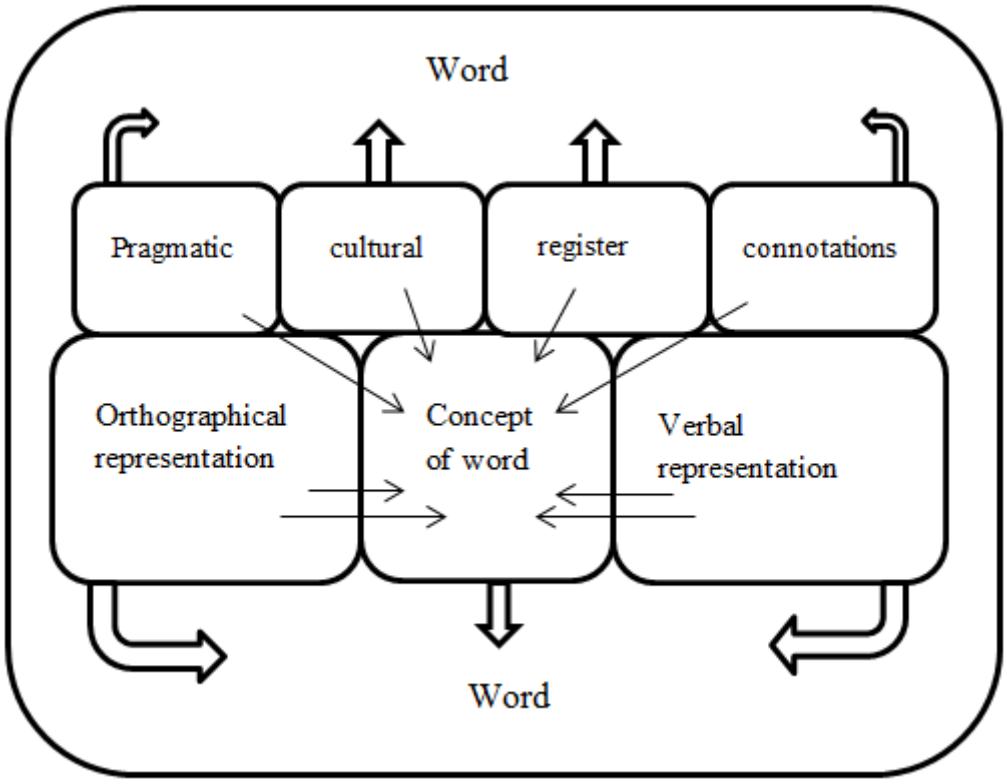

Figure 4. Visual representation of complex word knowledge (Buis, 2017)

In figure 4, the concept of word knowledge is depicted using several different components all linked to the concept of the word and the connection is shown by arrows internally and externally. The concept of the word is what it abstractly represents, while the word includes all the components including the orthographical, and the phonological representation. In terms of how de Saussure's been describing it, the signified is the word concept in the middle, while the signifier is the word at the bottom and the top (for clarity shown twice). Further components such as pragmatic, cultural, register and connotations may be regarded as additional features of a word. To illustrate this further, the words *chair* and *recliner* are synonyms in English. Even though they may differ slightly on the conceptual level, a chair is an object elevated off the floor upon which a person can sit, and a recliner is in this manner the same, however a chair is often an object with four legs, and a backrest: The concept of a recliner is a more comfortable, sturdier object, and often without four legs but with a single solid leg. The concepts of chair and recliner do not share all features on the word level. However, they do share the syntactic category of noun, several semantic features, and

the same pragmatic category. Complex knowledge of these words ensures that *recliner* is not selected when a speaker wants to make use of the concept of a chair, even though these words may be synonyms. Furthermore, the connotations of a chair and a recliner may differ per culture and age; young children may associate recliners with their grandparents, and in some speech communities recliners may not be as common as in others.

Thus, while several components of a word may fall in the same category and the basic concept of the word and synonym are the same (for example, chairs and recliners are both objects elevated from the floor to be sat upon), they should not be treated as equals. The current use of synonyms in vocabulary test questions is problematic because it only uses the similarities between words instead of also using the dissimilarities, which ensure that there are not two words in a language denoting two exactly similar concepts. Therefore, instead of focusing on the similarities, a better way to gauge in-depth knowledge of a word and its synonyms would be to enquire about what makes them different and in this manner test the knowledge which is acquired when synonyms are learned. One manner in which this may be done would be to use a context question in which the pragmatic knowledge of a word and its synonym is tested. For example, a context question could show a picture of a girl playing fetch with her dog. The question underneath the picture reads: ‘The dog returns the stick, what does the girl say? Pick the best option.

a) sufficient work

b) good job

c) poor job

d) adequately done

The only truly incorrect answer is option c, because it includes an antonym of good. However, the most suitable option in this context is option b, since option a and option d use a register which does not fit in the context wherein a girl is playing with her dog. This kind of question

may illustrates the difference between receptive knowledge and productive knowledge. Receptive knowledge could potentially result in the reasonable selection of options a, b, and d, while productive knowledge should only be able to produce option b in this specific context.

Thus, similarities are currently used when assessing synonym knowledge on vocabulary test items, even though the dissimilarities might be a better fit to assess knowledge beyond basic word knowledge such as pragmatic and cultural knowledge. Even though multiple-choice questions may use synonyms to assess the size of a learners' vocabulary, and whether the basic concept of synonymous words are known, this manner of testing is unable to evaluate other vocabulary knowledge, such as pragmatic knowledge. It must be noted that assessing word knowledge by using synonyms in a multiple-choice format is quite suitable for beginning language learners and learners of low abilities. Pragmatic competences, cultural sensitivity, and knowledge of register and word connotations are almost subconscious skills in L1 and L2, and require a proficient speaker. In a similar manner, pragmatic competences are not often learned in the classroom setting, but currently there are debates about whether teaching pragmatics should be practiced to facilitate easier foreign language communication (Fernandes, 2016; Ishihara & Cohen, 2014). When these competences are looked up within the CEFR (see appendix A, on page 80), it becomes clear that there is little to no mention of these competences or knowledge. However, even though pragmatic knowledge, cultural sensitivity and knowledge of register and connotations are not mentioned to a great extent within the CEFR, they facilitate communication in a foreign language and should be treated as equally important aspects in both vocabulary teaching and testing.

3.3 Vocabulary test question format

The relation between words and their synonyms, and some popular theories on vocabulary acquisition, have been thoroughly investigated, but other factors also play a crucial role in vocabulary tests. Certain vocabulary question item formats, which are used to estimate the size or the receptive knowledge of a language learner, may over- or underestimate the vocabulary knowledge and thus the size of the vocabulary. Depending on the question item format of the test items (e.g. multiple-choice questions), reading and writing abilities may be reliably tested in the foreign language. Thus, to be able to interpret the results in a reliable manner, it is important to investigate the manner in which vocabulary questions are formulated and to assess how these influence the test results.

The English IEP vocabulary test employs five different question formats. The multiple-choice method presents a question with three wrong answers and one correct answer. This method is able to test two words at once, namely the word which is asked in the question and the correct synonym provided in the answers. An example of a multiple-choice format and word-picture format as found in the English IEP vocabulary test is given in figure 5.

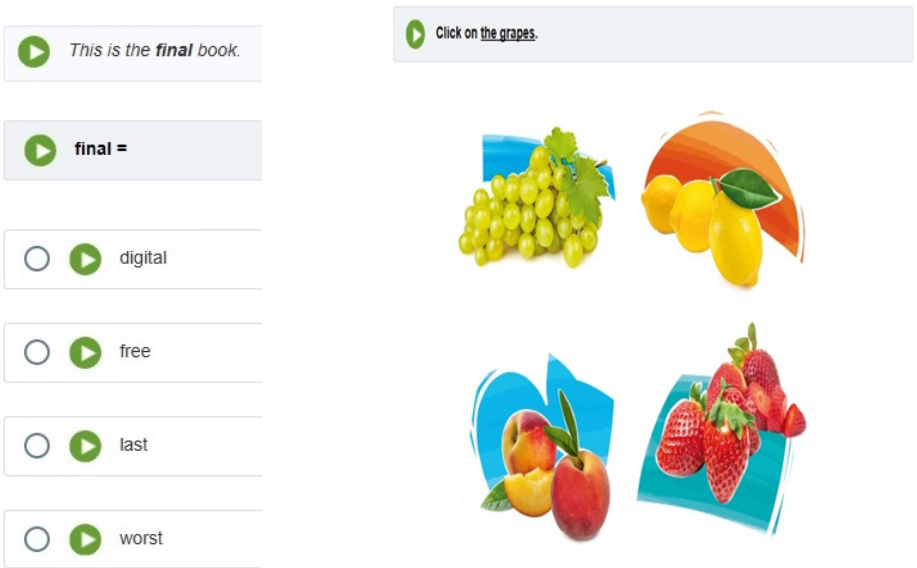

Figure 5. Multiple-choice and Word-Picture question format (English IEP vocabulary test, 2016).

As depicted in figure 5, a slight variation on the multiple-choice format is the word-picture format. This item is presented in the English IEP vocabulary test as instructions plus a word and four possible options underneath. This is a slight variation on the multiple-choice format, but instead of one synonym and three distractors, visual representations of the words are used. Yet another slight variation on the multiple-choice format is the picture-word format, in which a picture is shown accompanied by a questions with multiple-choice answers. An example is given in figure 6.

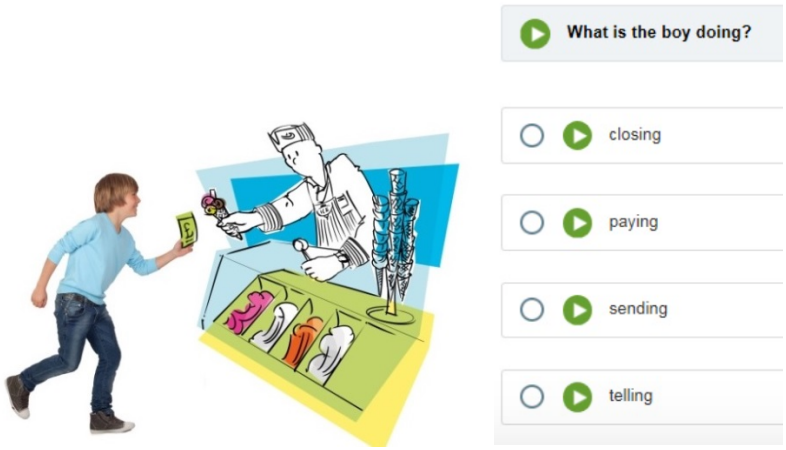

Figure 6. Picture-word question format (IEP, 2016).

Another form-meaning format is multiple matching, and this method presents two categories and several words which either belongs in the left or the right category. An example of a multiple matching format is given in figure 7.

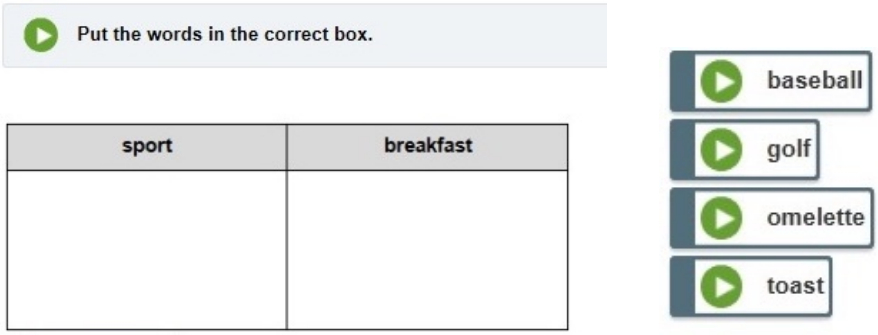

Figure 7. Multiple matching (IEP, 2016).

An alternative method of multiple matching, namely, mix and match is also used in the English IEP vocabulary test. Even though the underlying principle remains similar to the

multiple-choice format, there are several differences. A multiple-choice format enquires about one word and one correct synonym, and the other possible choice options function merely as distractors. In a multiple matching format 3 words are employed and 3 possible options are given, each option belonging to a specific word, resulting in a question format in which no distractors are used, this is illustrated in figure 8.

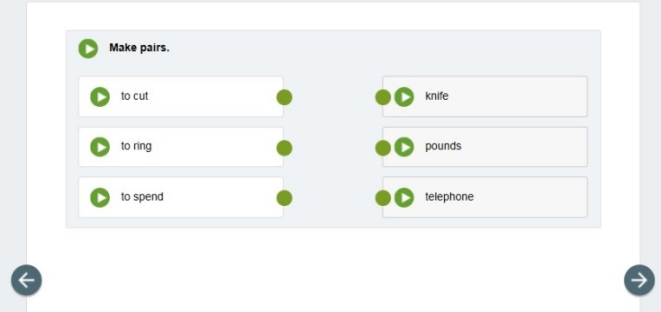

Figure 8. Mix and Match question format (IEP, 2016).

A correct answer on a vocabulary test item is often assumed to indicate that a word is known or learned. However, the problem with this assumption are the definitions used for the words *known* or *learned*. More often than not it is assumed that *known* or *learned* refers to the ability to have mastery of the word and be able to employ it in either reading, writing, speaking or listening. Nevertheless, this assumption is largely unsubstantiated for most vocabulary test formats (McLean, Kramer & Beglar, 2015). An item which has been answered correctly could also be interpreted as the learner knowing the register, collocations and derivations, but little to no information is available on how informative different formats can be about these aspects (Nation, 2001).

3.3.1 Method

Here, the formats in which the vocabulary words are presented on the English IEP vocabulary test will be investigated by comparing the different formats and their specific uses among each other. The most recent research towards predictability of vocabulary test items and the over and under projection of results will be discussed and presented, and compared to

the vocabulary test items found on the English IEP vocabulary test. The results will be presented and their implications will be discussed.

3.3.2 Results

There are several different formats in which test items may be presented. The most commonly known vocabulary item format is most likely the multiple-choice format. Other common vocabulary test formats include multiple matching, mix and match, and the picture-word/word-picture format. Kremmel and Smiths tested some of these formats to investigate what the results say about test takers' ability to employ words, specifically with respect to their reading ability (2016). The goal of their study was to understand the relationship between a correct answer on a vocabulary test item and receptive knowledge of the word in question given that a. A correct answer on a vocabulary test may be due to other factors than receptive or productive vocabulary knowledge, such as guessing, and b. A vocabulary test often uses several different test-item formats. The different question item formats were tested, and the knowledge of the words which occurred in the questions were afterwards confirmed using in depth interviews with the participants. In this manner, Kremmel and Smiths were able to measure how many words participants knew and how much certain question formats overestimated or underestimated vocabulary knowledge. It was found that some items overestimated or underestimated the vocabulary knowledge, and if the goal of the vocabulary test is to measure size, over- or underestimate the vocabulary size. One of their findings was that some items reflected vocabulary knowledge better, as to how many words a learner truly knows, than others (Kremmel & Schmitt, 2016, 388). The different vocabulary formats can be divided into two categories: form-meaning formats and form-recall formats. The first of these categories, the form-meaning format, makes use of, for example, the multiple-choice method and the multiple matching method. The second category, defined by Kremmel and Smiths as form-recall formats makes use of a cloze-format. However, these types of question item

formats do not occur in the English IEP vocabulary test and will therefore not be discussed further.

The multiple-choice and multiple matching test item formats are 'popular' and are often used because they have an apparent aptness for measuring individual words, are practical in use, are able to give a high sampling rate in a relatively short amount of time, and give 'objective' scores (Kremmel & Schmitt, 2016). However, one of the main criticisms concerning these test item formats is the guessing probability. In a number of studies, the possibilities for guessing the correct answer on a multiple-choice item has been examined, Kamimoto (2008), and in a similar manner Webb found there is roughly a 17 % chance of blindly guessing the correct response in a multiple matching format (2008). Stewart noted that when measuring receptive vocabulary, multiple-choice items are to be avoided as they inflate the vocabulary test scores due to the guessing probability which statistic correction formulas are unable to correct for due to the individual nature of the test taker (2014). However, this problem concerning multiple-choice also occurs with other question formats where multiple items to choose from are, or guess on, are presented, which may skew the results of whether a word is familiar, known, or neither.

Kremmel and Smith found that when investigating the multiple matching and multiple-choice item formats both these question item formats scored in a similar manner. The results of their study towards these two item formats measured in 99 participants are presented in table 3.

Table 3. Results study by Kremmel and Smith (2016)

	Multiple matching	Multiple-choice
Correct	54.9 %	56.7 %
Overestimation	22.2 %	20.3 %
Underestimation	3.3 %	2.4 %

Incorrect	19.6 %	20.7 %

According to Kremmel and Smith the multiple matching format has a 22.2 % chance on average for overestimation of receptive vocabulary knowledge, and an underestimation of 3.3 % (2016, p. 387). While the multiple-choice items showed an overestimation of 20.3 % and an underestimation of 2.4 %. Thus, while both options show an overestimation around 20%, this may be adjusted with scoring formulas. Although scoring formulas have been proposed to adjust the scores for guessing (e.g. correction for guessing formula (Huigbregtse, Admiraal & Meara, 2002)), these have been made under the assumption that the likelihood for guessing is the same for each participant and remains stable throughout the entire test (Stewart & White, 2010). However, the likelihood for guessing is dependent on factors such as vocabulary knowledge and how often test taking strategies (e.g. guessing or tactical choosing) are employed.

Item response theory (IRT) is a theory of testing based on the idea that not all items have the same difficulty level. Furthermore, it is a theory based on testing the individual performance on a test item, and their overall performance with regards to the ability that item was designed to measure. In IRT the 3-parameter logistic model by Birnbaum (1968) has been designed to link an individual's ability to item difficulty (Brown & Hudson, 2002). However, on tests such as the IEP using individual's ability to item difficulty is difficult because 'distractors' (other possible answer options) are chosen from the same CEFR level as the correct answer, and thus also come from the tested domain (Stewart & White, 2010). This results in possible inflations in test scores depending on the proportion of words known by the test taker.

The relationship between proportions of words known and increases in test scores due to guessing has been researched by Stewart and White (2010). While a priori this may seem similar to the study undertaken by Kremmel and Smith (2016), the difference lies in the fact

that Stewart and White did not focus on what certain vocabulary question item formats reveal about vocabulary knowledge, but focussed on the guessing aspects when it concerned a question with multiple options. They analysed the vocabulary level test[4], and how the knowledge of the test takers relates to their ability to guess items correctly. They wrote a programme in C++ to run a number of virtual stimulations of the VLT, keeping the variables as precise as needed. Furthermore, they tried to find a solution to several problems with scoring adjustment formulas, since these do not take into account the inconsistent guessing effect throughout the test and the predictability of how often an individual will guess an item compared to their overall abilities (Martin, del Pino, & De Boeck, 2006). When a multiple-choice question is posed, of which the format is 1 question with 4 possible choices, and the test taker has no knowledge of any of the test items, they must guess three times resulting in a correct guessing probability of ¼ time for each individual guess. However, when the test takers know none of the words in question, but does know two of the distractor words, they still have to guess once, but due to an elimination process, they now have a ½ change of guessing correctly (Stewart & White, 2010). The results of their study showed a steady increase of approximately 16-17 points until over 60 % of the words are known.

However, they note that the guessing effect was estimated at intervals of 5 points, which may lead to individual results (increases due to guessing) which are higher or lower than predicted in their study. The results show that there is a ceiling effect for guessing, thus for language learners who know over 80 % of the words on a vocabulary test it will be impossible to see an increase from guessing over 16 %. While for the lower ability language learners, the possibility to guess correctly rises gradually until 80 % of the words are known.

Knowledge of the tested domain increases guessing ability, while increased knowledge of the tested domain also increases chances that the test-taker knows the correct

[4] The VLT is a controlled production test.

answer. The relationship between guessing and knowledge in the vocabulary domain is a complicated one and that it remains difficult to devise a formula which can take these variables into account.

An alternative to using formulas to adjust for guessing has been proposed, which is the 'I don't know' option. When faced with a multiple-choice question and the answer is unknown, there is in a traditional format (one question, four or six possible answers) nothing else to do than make a forced guess on one of the four or six possibilities. A solution which might reduce guessing is a fifth or seventh answer option, namely: I don't know (IDK). Researchers have reported reduced guessing and improved estimates of reliability (Lucovich, 2014; Zhang, 2013). The IDK option, however, would be most appropriate when the test is not taken for a grade, but rather to estimate the vocabulary size of a pupil; otherwise the pupil might guess to possibly inflate their grade.

Some factors may make the IDK option not a suitable solution. Since some learners use the IDK option more often than other learners (Bennet & Stoeckel, 2012; Zhang, 2013). Thus, learners with a similar vocabulary could achieve vastly different test scores (Stoeckel, Bennet & McLean, 2016). While the question remains "what is being tested with vocabulary tests?", the inclusion of the IDK option has been investigated by Stoeckel, Bennet and McLean (2016). They designed a study using 1,000 computer-generated learners who were simulated to be normally distributed across a wide range of vocabulary knowledge (Stoeckel, Bennet & McLean 2016). To account for the difference in learners, they designed several conditions in which all known items were answered correctly. In the first condition, all unknown items were answered incorrectly, representing actual ability (Stoeckel, Bennet & McLean 2016). In the second condition, learners guessed on all items they did not know representing a normal vocabulary test; in all other conditions a portion of the learners always guessed when they did not know the word and another part always answered with IDK

(Stoeckel, Bennet & McLean 2016). The reported results show that unless 93 % of all participants use the IDK answer instead of guessing, the ordering of participant per ability is poorer than when IDK is not a viable option on a vocabulary test. The spread of scores is reduced, making the test's ability to discriminate lower, thus reducing reliability (Ary et al., 2013). Furthermore, the results show that the link between actual knowledge and test scores becomes less reliable, due to partial use of the IDK answer, and partial guessing; no formula can be constructed to adjust for this. However, the computer simulation could not account for partial use of the IDK answer and partial knowledge. Real learners often do use partial knowledge of a word or test strategies to correctly answer vocabulary test items and vary the use of the IDK answer considerably (Gyllstad, Vilkaite, & Schmitt, 2015; Lucovich, 2014; Zhang, 2013). Nevertheless, the use of an IDK answer on multiple-choice vocabulary test items may not be the solution needed to reduce the possibility of guessing as it diminishes the validity by weakening the link of person-ability to item difficulty (Stoeckel, Bennet, & McLean 2016).

Thus, depending on the amount of knowledge a test-taker has when taking a vocabulary test, until 80 % of the words on the test are known the test scores do not resemble true knowledge of the test-taker. Rather the results are inflated by guessing. While no knowledge results in a guessing probability of ¼ (or 1/6 depending on the number of possible answers), these chances become higher the more knowledge a test-taker has because they are able to eliminate some of the wrong answers and more easily pick the correct answer.

The individual nature of vocabulary knowledge ensures that it remains difficult to devise a single formula which will be able to consistently adjust test scores to reflect true test taker knowledge. While the multiple-choice and multiple matching format may increase the test scores for part of the test-taker group, using formulas to adjust the test scores seems ill-advised due to the fact that increased knowledge leads to better guessing probabilities but also

to a better probability of knowing the correct answer. Furthermore, not all test-takers will behave in a similar manner due to individual competences and knowledge. Overall, standardized vocabulary test can hardly give an accurate estimation for an individual's vocabulary size, but may give a valid estimation of receptive and productive vocabulary knowledge for individual test takers.

From the study by Kremmel and Smith it became clear that both multiple-choice and multiple matching question format over-project and under-project the results. However, the English IEP vocabulary test has three other question item formats which have not yet been investigated thoroughly, so no information about the relation between a correct answer and actual vocabulary knowledge is available on mix and match, picture-word, and the word-picture question item format. The different distribution of the item formats of the English IEP vocabulary test is presented in table 4.

Table 4. English vocabulary IEP template

	Multiple matching	Multiple-choice	Picture-word	Word-picture	Mix and match
Total	17	9	2	8	9

Table 4 demonstrates that the multiple matching question item format is represented the most with 17 questions and the multiple-choice question format is represented with 9 items. This means that 17 items may over project the results for 22.2 %, while 9 items may over project them for 20.3 %. These items may in turn under project them for 2.4 % and 3.5 % respectively. As stated earlier, the current goal of the IEP vocabulary test is to give an indication about the CEFR level which the pupil has achieved. This CEFR level is used to predict the level of reading and listening competences in the foreign language. This would indicate that the IEP vocabulary test has as a second objective of testing receptive vocabulary knowledge (i.e. knowledge necessary for reading and listening). However, the formats used in the IEP vocabulary test may not be suitable for this purpose. To be able to read and listen

fluently, a quick recognition of the word forms, following an automatic retrieval of the meaning is necessary so that cognitive resources can be applied to construct meaning from the text (Grabe, 2009). The skill of receptive vocabulary would therefore include knowledge of the word to the *meaning recall level* (Schmitt, 2010). In a natural reading situation, a word must be recognized without any help or options to choose from for (unknown) words in the text (Nation & Webb, 2011). Vocabulary items which are presented in the form of multiple-choice or multiple matching are incongruent with the skills necessary for receptive vocabulary knowledge such as reading and listening and are therefore named alternative receptive skills. These formats test words on a recognition recall level where options are given and must be selected from. While these skills are somewhat similar to reading and listening skills, the differences lie in the fact that alternative options are given which may be used to infer the correct answer. In all the alternative receptive dimensions of vocabulary testing, multiple-choice, multiple matching, and mix and match, there is either a synonym to match (multiple-choice and multiple matching) or words that need to be put into a corresponding category (mix and match). These options to choose from when retrieving the definition of a word would not be available in a non-testing situation, which has led to my proposal of a new dimension of alternative receptive vocabulary competence.

The picture-word / word-picture format used in the English IEP vocabulary test, however, is an example of a question format which is semi-congruent with receptive and productive vocabulary skills. When a word is presented after which choice of the correct picture is asked, the situation becomes a bit more like a natural reading situation, after all a word is read and the corresponding concept (depicted visually) must be chosen. The other way around, in a which a picture is shown and a possible option of answers exist out of a corresponding word for the picture and some distractors, this may measure the least receptive ability – a picture is shown after which the correct corresponding word must be chosen. This

skill mostly resembles productive skills wherein an individual may want to use a word to describe a concept (or picture) and chooses a congruent word to employ. However, since the test taker does not in fact produce anything, this skill is called an alternative productive skill, while a classic productive skill would be a cloze test in which a prompt is given, often in the form of a sentence with an __ for the word which the test taker must produce. An overview of different vocabulary test item formats and their corresponding vocabulary dimensions are given in table 5.

Table 5. Vocabulary dimensions per question item format

Vocabulary dimension	**Question item format**
Classic receptive	Word-picture
Classic productive	Cloze test
Alternative receptive	Multiple-choice, Multiple matching, Mix and Match
Alternative productive	Picture-word

In conclusion, it appears that the English IEP vocabulary test tests both alternative as well as classic receptive skills, while in fact it only tests alternative productive skills. It may thus be argued that the results from the English IEP vocabulary test represent skills which are overall most closely related to receptive language competences, but the results also may not necessarily represent the totality of real life language competences.

3.4 Predictive relationship between vocabulary test and receptive and productive competences

The predictive relationship between having a large vocabulary and being a proficient reader and writer has been established (Beglar, 2010, Meara, 1996, Meara & Buxton, 1987, Meara & Jones, 1988, Nation, 1990). This relation seems intrinsic as "learners with big vocabularies are more proficient in an array of language skills than learners with small vocabularies" (Meara, 1996). Nonetheless, a large vocabulary is needed to produce an array of sentences rapidly or read and comprehend differing words quickly. No widely accepted theory has been yet defined as to what leads to lexical competence and no research towards this has been done. It remains important to ensure that a vocabulary test will be able to give the most reliable and valid results for both the educator and the test taker. The more a vocabulary test item format produces reliable answers (e.g. the answers represent knowledge of the test taker), the higher the 'predictability factor' of a test becomes. Thus, the more reliable a vocabulary test is, the easier it becomes for the test to give an indication about language competences based on the results of the vocabulary test. However, until the relationship is further investigated and established, vocabulary test scores should not be used as a leading factor in measuring language ability.

4. The English IEP vocabulary test and possible future improvements

The research done here on an array of topics from synonyms to question item format will be compared and contrasted with the current English IEP vocabulary test from Bureau ICE in this chapter. A short overview per topic of the results will be given, followed by a discussion of the results and the implications these have. Possible recommendations will be given to improve the English IEP vocabulary test where this may seem necessary.

4.1 Synonyms

The nature of the multiple-choice question item format in vocabulary tests has been examined with regards to the nature of synonyms. More often than not, synonyms are used to measure word knowledge of both the word in question and the correlating correct synonym. However, several issues arise when measuring vocabulary knowledge in this manner. Even though a synonym and its corresponding word share some correlating resemblances on a basic concept level, enquiring about vocabulary knowledge in this manner is only able to confirm knowledge on a basic concept level. The components which are unable to be measured in this manner are pragmatic competences, knowledge of register, connotations, and cultural sensitivity knowledge. Thus, while the similarities between the word in the vocabulary question and the corresponding synonym is an appropriate way to measure basic word knowledge, it fails to measure deeper word knowledge. This is problematic because these components of word knowledge are as necessary as word concept knowledge to become a proficient competent communicator in a language. To utilize not only the similarities between a word and its synonym, this study proposes a new vocabulary question format enquiring about these different components of a word, such as pragmatic competence which may need to take a more central role in foreign language teaching to ensure all language learners are competent communicators who are able to make use of all a word's knowledge levels.

While vocabulary test results are interpreted often by the teacher and the pupil as knowledge the test taker has, there remains little to no clarity as to what the test taker precisely knows about these words. The results here imply that when measuring vocabulary knowledge in this manner used in current tests only the basic concept similarities between a word and a synonym are measured, and thus current tests do not account for possibility that not all components of word knowledge have been acquired yet by the test taker. This study proposes a new idea on how to use synonyms and words to measure vocabulary knowledge by

utilizing the components of word knowledge such as knowledge of pragmatic and register, which may be used for advanced language learners. Furthermore, information about what a vocabulary test measures precisely should be more easily available. More information about the nature of vocabulary tests ensures the teacher and the language learner are able to interpret their results in an accurate manner. A better representation of the nature of the acquired vocabulary knowledge provides both the teacher and the language learner with a more complete overview of the components of word knowledge which are already acquired and the components which require more learning.

In conclusion, using synonyms to measure basic word concept knowledge is appropriate, while the other dimensions of word knowledge (pragmatic competences, knowledge of register, connotations, and cultural sensitivity) may be used to measure a higher level of word knowledge. An example of the question method regarding pragmatic knowledge and other vocabulary knowledge is given in 3.1, on page 31. Furthermore, teaching and testing of these other components of word knowledge should be treated as equally important aspects of vocabulary teaching and testing.

4.2 Vocabulary Test Question Format

The role of question format in vocabulary testing has just recently started to become the topic of research. While no information was available on the over and underestimation on mix and match, word-picture, picture-word question item formats, these formats may also overestimate vocabulary knowledge due to the multiple option format they employ. This may lead to a very high likelihood that the question format items of the English IEP vocabulary test may overestimate the vocabulary knowledge of the test takers. However, the English IEP vocabulary test's results may be used as a formative assessment rather than as summative assessment. The score of the overall English IEP test are currently used to give an indication

of the CEFR level of the pupil, which may be used to structure English secondary education to fit the pupils needs better.

Several solutions to the overestimation of vocabulary test questions with multiple answers have been investigated such as an IDK option, or an adjusting scoring formula which may be applied afterwards. However, none of these options appeared viable as both guessing behaviour as using the IDK option are largely dependent on individual abilities and behaviour. Instead, this study proposes that the teacher should be aware that the vocabulary test results may not represent true knowledge of the words tested and that their remains a probability for correctly guessing the correct answer.

The knowledge of what vocabulary test results represent is important in interpreting test results and in structuring education in a manner to facilitate growth instead of stagnation. Furthermore, when a certain question item format is shown to overestimate vocabulary knowledge this format should not be used as much in future vocabulary tests. New question item formats may be developed, or investigated with regard given to how much they over or underestimate vocabulary knowledge. The question item formats mix and match, word-picture, and picture-word should be investigated as to how a correct answer relates to actual vocabulary knowledge to be able to interpret the scores of the English IEP vocabulary test in a reliable manner.

4.6 Predictability Properties

There have been several studies that have found a relationship between vocabulary size and language competence. However, this relationship between vocabulary size and language competence appears to influence each other, since a larger vocabulary leads to a more rapid production and understanding of sentences (either written, or spoken), while being proficient in both receptive and productive tasks leads to a faster acquisition of new

vocabulary words. Thus while vocabulary knowledge plays an important role in how proficient a language learner is (or becomes), other factors such as personal aptitude, attitude and socioeconomic status may also play an important role and may appear as predictive as vocabulary knowledge for language competences. However, isolating one factor and attributing individual language competence to this may be dangerous as the relationship might not be so clear cut. When a pupil has a weak performance on a vocabulary test, there may be other factors contributing to this than just the lack of being a proficient language user. Even language learners with a small vocabulary may be competent and proficient language users in a foreign language. If a language learner is only judged based on their vocabulary, they may be categorized as a low ability language learner while they may just need to expand their vocabulary instead of improve their overall language abilities.

In the English IEP test, the results from the first part (vocabulary) are used to give a starting point for the second part of the test, which can be either A1/A2 or A2/B1. Thus a pupil may unnecessarily be placed in the low category at the listening part of the test, while their listening skills are competent enough to achieve a higher score. However, after the listening part of the test the pupil will be, based on the scores achieved in the vocabulary test and the listening test, redirected to the appropriate level for reading and listening; this is on CEFR level A1, A1/A2, A2, A2/B1 or B1. The English IEP vocabulary test uses the predictability scores of the test, but there is also room for the pupil to exceed the previously set expectations.

Thus, until the relationship between vocabulary knowledge is further investigated and established, vocabulary test scores should not be used as a leading factor in predicting language ability as is currently done in the English IEP vocabulary test.

5. Conclusion / discussion

In this thesis, the conclusive English IEP vocabulary test for Dutch pupils has been discussed. The research focused on answering the research question: "What competences are being tested with vocabulary test items?" Furthermore, the sub-questions of this research focused on the predictive relationship between vocabulary test items and language competence for Dutch English foreign language learning children and their future English competences. The research also focused on whether vocabulary test items are reliable and valid, and whether reliability and validity differ between the vocabulary question item formats. Furthermore, it was investigated whether improvements should be made to the English IEP vocabulary test regarding validity, and reliability. First, the answers to the sub-questions will be summarized, after which the research question will be answered and discussed. In addition, suggestions for future research will be made and the limitations of this study will be stated.

1) What are the predictability properties of vocabulary test items for Dutch English foreign language learning children and their future English competences?

While the positive relationship between vocabulary knowledge and language competences has been empirically proven, the relationship may be somewhat exaggerated. While vocabulary evidently influences receptive and productive language competences, it should not be the leading factor in determining language competences. Other factors such as personal aptitude, attitude and socioeconomic status also play an important role and may be as predictive in a similar manner as vocabulary knowledge is for predicting language competences. Thus, while there is indeed a relationship between vocabulary test results and (future) language competences, other factors should not be disregarded when determining language competences as a small vocabulary may not necessarily be an indication of low language abilities and vice-versa.

2) How reliable are vocabulary test items?

Several vocabulary test item formats have been investigated by Kremmel and Smith (2016), and from these results it became clear that certain question item formats overestimated and underestimated vocabulary knowledge. Other studies (Stewart & White, 2010; Webb, 2008; Kamimoto, 2008) have also shown that vocabulary test items with multiple options give rise to a possibility of guessing the correct answer. With no viable alternative to vocabulary test items with multiple options, the results of vocabulary tests may not be as reliable and valid as necessary to interpret the knowledge of the test taker as reflective of the test results (Kremmel & Smith, 2016).

3) How does the reliability differ among vocabulary test item formats?

It was shown that the differences between multiple-choice and multiple matching question item formats differed, with a corresponding 20.3 % and 22.2 % overestimation, and a small amount of underestimation. Unfortunately, due to the scope of this thesis, the vocabulary test item formats used in the English IEP vocabulary test were unable to be tested in this manner, and research on other test formats needs to be done, such as multiple matching, picture word, and word picture. It is likely, however, that these question item formats will all result in different grades when measuring vocabulary knowledge due to the multiple option formats. However, the smaller the probability becomes to guess the correct answer, the more reliable the vocabulary test items become.

4) In what way can improvements be made to the English IEP vocabulary test regarding reliability and predictability properties?

Several suggestions have been given to improve the English IEP vocabulary test, such as using other dimensions of vocabulary knowledge to test higher CEFR level abilities. The use of synonyms to test word knowledge as is currently done in the English IEP vocabulary test may be suitable for A1-B1 level, but other dimensions such as pragmatic use, formality,

culturally awareness of sensitive words, and connotations should also be tested at higher levels. Doing this would ensure that vocabulary tests test language learners knowledge of the basic concept of words but also how to employ these words.

Furthermore, recommendations for future English IEP vocabulary testing include refraining from using adjustment formulas for guessing, since it is nearly impossible to predict guessing behaviour for a group. Guessing behaviour will always differ among individuals based on vocabulary knowledge and test taking strategies. The results from this study also suggest refraining from continuing to presume to test vocabulary size, since the extrapolation of the vocabulary tests' results will lead to unreliable estimations of the size of the vocabulary.

This research has been conducted to answer the following research question: What competences are being tested with vocabulary test items? The focus was specifically on vocabulary test items from the English IEP vocabulary test. It became evident that a definite answer would not be available, since not enough research on vocabulary test items has been done. The established competences which vocabulary test items measure (receptive and productive competences) have been extended to include often used vocabulary test items which do not appear to fall into the classic categories. The new addition to the categories which was proposed in this thesis are alternative receptive competences and alternative productive competences. While the differences between the alternative and the classic vocabulary categories are apparent, the similarities are enough to categorize these in the alternative dimension of the respective classic dimension. However, the differences which include the possibility to choose from several answers, use a strategy to arrive at the correct answer such as guessing or inferring, and using synonym knowledge to choose the correct answer ensure that these testing methods do not fall into the classic receptive category. Thus it has been proposed here that the multiple-choice, the multiple matching, and the mix and

match item format will fall into the proposed category of alternative receptive competences. A manner to measure alternative productive competences includes the picture-word task, in which a test taker is still aided towards the correct answer and able to use test strategies in a similar manner as with the alternative receptive competences. However, due to the possibility to guess the correct answers on multiple-choice questions, the answers may not represent true vocabulary knowledge or language competences. Additionally, the multiple matching, and mix and match question item format may also be prone to guessing. However, the distractors are not quite distractors in the classic notion as they are in the multiple-choice format, since these also need to be matched to either a synonym (mix and match) or to a category (multiple matching). The precise nature of the relationship between vocabulary knowledge and competences remain unclear, and more research towards this subject will be necessary to interpret vocabulary test results in a viable manner in the future.

The limitations of the current study were that this thesis presents were unable to include quantitative research necessary for conclusive evidence. a mix of already existing research and new research done by this author to give an extensive overview of the English IEP vocabulary test. Some of the already existing research done by Kremmel and Smith (2016) may be replicated in the future to give an even more accurate picture of the situation regarding the English IEP vocabulary test and lead to more conclusive answers regarding what competences vocabulary tests ‘test’.

Further research should focus on the nature of the relationship between vocabulary knowledge and vocabulary test items. To be able to interpret test results in a valid and reliable way, and the nature of what competences are being tested needs to be clear. The current information on this subject is too limited to interpret test results in a satisfying manner.

Bibliography

Adams, A. M., & Gathercole, S. E. (1996). Phonological working memory and spoken language development in young children. *Quarterly Journal of Experimental Psychology A, 49,* 216-233.

Adams, A. M., & Gathercole, S. E. (2000). Limitations in working memory: implications for language development. *International journal of Language Communication Disorders, 35,* 95-116.

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of memory and language, 53*(1), 60-80.

Ary, D., Jacobs, L., Sorensen, C., & Walker, D. (2013). *Introduction to research in education* (9th ed.). Belmont, CA: Wadsworth Cengage Learning.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), The psychology of learning and motivation: Advances in research and theory (pp. 89–195). New York, NY: Academic Press.

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature reviews neuroscience*, *4*(10), 829-839.

Baddely, A. D., Gathercole, S. E., & Papagano, C. (1998). The phonological loop as a language learning device. *Psychological review, 105*, 158-173.

Baddely, A. D., & Hitch, G. J. (1974). Working memory. *Psychology of learning and motivation, 8,* 47-89.

Baddely, A. D., & Logie, R. H. (1999). Working memory: The multiple component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). New York: Cambridge University Press.

Beck, I. L., McKeown, M. G., & Kucan, L. (2008). *Creating robust vocabulary: Frequently asked questions and extended examples* (10th ed.). New York, NY: Guilford Press.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118.

Bennet, P., & Stoeckel, T. (2012). Variations in format and willingness to skip items in a multiple-choice vocabulary test. *Vocabulary Education and Research Bulletin*, *1*(2), 2-3.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University press.

Browne, Chihi, Culligan (2007) Measuring Vocabulary size via Online Technology. Lexxica.

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment.* Retrieved from https://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37-54.

De Groot, A. M. B. (1993). Word-type effects in bilingual processing tasks: Support for mixed-representational system. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 27–51). Amsterdam: John Benjamins Publishing Company.

De Groot, A.M.B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*(1), 1-56.

De Groot, A.M..B. (2011). *Bilingual cognition: An introduction*. New York: Psychology Press.

De Saussure, F. (1916) *Cours de linguistique générale*. Bally, C. & Sechehaye, A. (Ed). Paris: Payot.

Dolch, E. W. (1936). A basic sight vocabulary. *The Elementary School Journal*, *36*(6), 456-460. Retrieved from: http://gemini.es.brevard.k12.fl.us/sheppard/reading/dolch.html

Ehri, L. C. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In Gough, P., Ehri, L., & Treiman, R., (Eds.), *Reading acquisition* (pp. 107–143). Hillsdale, NJ: Erlbaum.

Ehri, L. C., & Rosenthal, J. (2007). Spelling of words: A neglected facilitator of vocabulary learning. *Journal of Literacy Research, 3*9(4), 389–409.

Nuffic (2015) *Vroeg vreemdetalenonderwijs Engels. Visiedocument.* Retrieved from https://www.nuffic.nl/publicaties/vind-een-publicatie/vroeg-vreemdetalenonderwijs-engels-visiedocument.pdf

ERK. (N.D.). Europees referentiekader talen. FAQ's. question 6. Retrieved from www.erk.nl/ouders/FAQ/

Falchikov, N. (2005). *Improving assessment through student involvment: Practical solutions for aiding learning in highe rand futher education.* London: Routledge Falmer.

Fernandes, A. C. (2016). *Gender differential item functioning on English as a foreign language pragmatic competence test: Implications for English assessment policy in China* (Doctoral dissertation, Niagara University). Retrieved from ProQuest. (10127239)

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics, 19,* 463-487.

Gathercole, S. E., & Alloway, T. P. (2008).Working memory and classroom learning. In Thurman, K., & K. Fiorello, K. (Eds.), *Cognitive development in K-3 classroom learning: Research applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gathercole, S. E., & Baddeley, A. D. (1993). *Essays in cognitive psychology. Working memory and language*. Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Gerritsen, M., Korzilius, H., van Meurs, F., 7 Gijsbers, I. (2000). English in Dutch Commercials: not understood and Not appreciated. Journal of Advertising Research, *40*(2), 17-34.

Gernsbacher, M.A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity, and orthography, concreteness, and polysemy. *Journal of Experimental Psychology*: *General*, *113*(2), 256-281.

Geurts, B. & Hemker, B. (2013). Balans van het Engels aan het eind van de basisschool 4; uitkomsten van de vierde peiling in 2012. Retrieved from http://www.cito.nl/~/media/cito_nl/Files/Onderzoek%20en%20wetenschap/ppon/cito_ppon_balans_52.ashx

Grabe, W. (2009). *Reading in a second language*. Cambridge, UK: Cambridge University Press.

Gyllstad, H., Vilkaite, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *International Journal of Applied Linguistics*, 166, 278-306. Doi:10.1075/itl.166.2.04gyl

Hattie, J.A.C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hayek, F.A. (2014). The Vienna Circle. In Meredith, P. (Eds.), *Instruments of Communication: An Essay on Scientific Writing*, pp. 89. Elsevier.

Hue, C. W., & Erickson, J. R. (1988). Short-term memory for Chinese characters and radicals. *Memory & cognition*, *16*(3), 196-205.

Huigbregtse, I., Admiraal, W. & Meara, P. (2002). Scores on a yes-no vocabulary test: correction for guessing and response style. *Language testing, 19*, 227-245.

Ishihara, N., & Cohen, A. D. (2014). *Teaching and learning pragmatics: Where language and culture meet*. Routledge. Doi:10.1111/j.1540-4781.2012.01348.x

Kamimoto, T. (2008) *Guessing and vocabulary tests: Looking at the Vocabulary Levels Test.* Paper

presented at the 41 Annual BAAL Conference, Swansea, UK.

Kilgarriff, A. (N.D.) Lemmatized BNC frequency list [Data file]. Retrieved from http://www.kilgarriff.co.uk/BNC_lists/lemma.num

Kremmel, B., & Schmitt, N. (2016) Interpreting Vocabulary Test Scores: What Do Various Item Formats Tell Us About Learners' Ability to Employ Words?. *Language Assessment Quarterly, 13*(4), 377-392. Doi:http://dx.doi.org/10.1080/15434303.2016.1237516

Laufer, B., & Nation, I.S.P. (1995) Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Li, H., Zhang, J., Ehri, L., Chen, Y., Ruan, X., & Dong, Q. (2016). The role of orthography in oral vocabulary learning in Chinese children. *Reading and Writing*, *29*(7), 1363-1381.

Logie, R. H. (2014). *Visuo-spatial working memory*. Psychology Press.

Lotto, L., & De Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, *48,* 31–69.

Lucovich, D. (2014). Adding *I don't know* to the Vocabulary Size Test. In N. Sonda & A. Krause (Eds.), JALT2013 *Conference Proceedings.* Tokyo: JALT. Retrieved from httP://jalt-publications.org/proceedings/article/4049-adding-i-don%E2%80%99t-know-vocabulary-size-test

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40,* 230–262.

Martin, E., del Pino, G., & De Boeck, P. (2006). IRT Models for Ability-Based Guessing. *Applied*

psychological Measurement, *30,* 183-203.

Masoura, E. V., & Gathercole, S.E. (2005). Phonological short-term memory skills and new word learning in young greek children. *Memory, 13,* 422-429.

Marzano, R. J. (2013). Resilience and Learning. Art and Science of Teaching/Cognitive Verbs and the Common Core. *Educational Leadership*, *71*(1), 78-79.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer and J. Williams (Eds.) *Competence and performance in language learning*. Cambridge: Cambridge University Press. pp. 35–53.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*(2), 142–154.

Meara, P., & T. Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System, 28*(1), 19–30.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In Grunwell, P. (Eds.) *Applied linguistics in society*. pp. 80–87.

Mclean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary test. *Language Teaching Research, 19*(6), 741-760. doi: 10.1177/1362168814567889

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Bristol, UK: Multilingual matters.

Nation, I. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge, UK: Heinle and Heinle.

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: broader language skills contribute to the development of reading. *Journal of Research in Reading*, *27*, 342-356.

Nation, I. S. P., & Webb, S. (2011). Researching vocabulary. Boston, MA: Heinle-Cengage ELT.

Neufeld, S. & Eldridge, J. (2009). LexiCLIL: A lexical Syllabus for the common European Framework for English. N.P.: LexiTronics.

Nizonkiza, D. (2016). First-year university students' receptive and productive use of academic vocabulary. *Stellenbosch Papers in Linguistics*, *45*(1), 169-187.

Paradis, M. (1997). The cognitive neuropsychology of bilingualism. In A. M. B. De Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives*, pp. 331–354. Mahwah, NJ: Lawrence Erlbaum.

Pavičić Takač, V. (2008). Vocabulary learning strategies and foreign language acquisition (27th ed.). Clevedon, UK: Multilingual Matters.

Paivio, A. (1991). Mental representations in bilinguals. In A.G. Reynolds (Eds.), *Bilingualism, multiculturalism, and second language learning: The McGill Conference in Honor of Wallace E. Lambert,* pp. 113-126. Hillsdale, NJ: Lawrence Erlbaum.

Paivio, A. (2007) *Mind and its evolution: A dual Coding theoretical approach.* Mahwah, NJ: Lawrenec Erlbaum.

Paivio, A., & Lambert, W. (1981). Dual coding and bilingual memory. *Journal of Verbal Learning and Verbal Behaviour. 20*, 532-539.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing 10*(3), 355-371.

Rastle, K., McCormick, S. F., Bayliss, L., & Davis, C. J. (2011). Orthography influences the perception and production of speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1588.

Ricketts, J., Bishop, D. V., & Nation, K. (2009). Orthographic facilitation in oral vocabulary acquisition. *The Quarterly Journal of Experimental Psychology*, *62*(10), 1948-1966.

Schmitt, N. (2000). *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave Macmillan.

Slo (2015). Moderne vreemde talen Vakspecifieke trendanalyse. Retrieved from http://downloads.slo.nl/Repository/moderne-vreemde-talen-vakspecifieketrendanalyse-2015.pdf

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139-152.

Stahl, S.A. & Nagy, W.E. (2006). *Teaching word meaning*. Mahwah, NJ: Erlbaum.

Stemach, G. & Williams, W. (1988). *WordExpress: The first 2500 words of spoken English*. Novato, CA: Academic Therapy Publications.

Stewart, J., & White, D. A. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, *45*(2), 370-380.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly, 11*(3), 271-282. Doi:10.1080/15434303.2014.922977

Stoeckel, T., Bennett, P., & Mclean, S. (2016). Is “I Don't Know” a Viable Answer Choice on the

Vocabulary Size Test?. *TESOL Quarterly*, *50*(4), 965-975.

Toetsbesluit PO. (2014). Retrieved from: https://www.rijksoverheid.nl/onderwerpen/toelating-middelbare-school/documenten/besluiten/2014/01/20/toetsbesluit-po accessed on: 12-04-2017

Verhagen, J., Messer, M.H., & Leseman, P.P.M. (2015). Phonological memory and the acquisition of grammar in child L2 learners. *Language Learning, 65,* 417-448.

Verspoor, M. (2010). Binnen- en buitenschools taalcontact en het leren van Engels. *Levende Talen Tijdschrif, 11*(4,)14-33.

Vulchanova, M., Foyn, C. H., Nilsen, R. A., & Sigmundsson, H. (2014). Links between phonological memory, first language competence and second language competence in 10-year-old children. *Learning and Individual Differences, 35,* 87-95.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition, 30,*79-95.doi:10.1017/S0272263108080042

Webb, S. A. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review*, *65*(3), 441-470.

Wolford, G., & Hollingsworth, S. (1974). Lateral masking in visual information processing. *Attention, Perception, & Psychophysics*, *16*(2), 315-320.

Zhang, X. (2013). The I don't know option in the Vocabulary Size Test. *TESOL Quarterly, 47,* 790-811. Doi:10.1002/tesq.98

Appendix A. CEFR frameworks' can do- statements

Table 6. (Council of Europe, 2001, p. 26-29)

Proficiency Level	writing	reading	listening
A.1	Can understand basic instructions or take part in a basic factual conversation on a predictable topic.	Can understand basic notices, instructions or information.	Can complete basic forms, and write notes including times, dates and places.
A.2	Can express simple opinions or requirements in a familiar context.	Can understand straightforward information within a known area, such as on products and signs and simple textbooks or reports on familiar matters.	Can complete forms and write short simple letters or postcards related to personal information.
B.1	Can express opinions on abstract/cultural matters in a limited way or offer advice within a known area, and understand instructions or public announcements.	Can understand routine information and articles, and the general meaning of non-routine information within a familiar area.	Can write letters or make notes on familiar or predictable matters.
B.2	Can follow or give a talk on a familiar topic or keep up a conversation on a fairly wide range of topics.	Can scan texts for relevant information, and understand detailed instructions or advice.	Can make notes while someone is talking or write a letter including non-standard requests.
C.1	Can contribute effectively to meetings and seminars within own area of work or keep up a casual conversation with a good degree of fluency, coping with abstract expressions.	Can read quickly enough to cope with an academic course, to read the media for information or to understand non-standard correspondence.	Can prepare/draft professional correspondence, take reasonably accurate notes in meetings or write an essay which shows an ability to communicate.
C.2	Can advise on or talk about complex or sensitive issues, understanding colloquial references and dealing confidently with hostile questions.	Can understand documents, correspondence and reports, including the finer points of complex texts.	Can write letters on any subject and full notes of meetings or seminars with good expression and accuracy.